World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3040

JOINT SEGMENTATION MODEL FOR CRACKS AND JOINTS BASED ON Deeplabv3+

Fang Wang
School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China.
Corresponding Email: wangfang05092022@163.com

Abstract: This paper addresses the problem of low segmentation accuracy of cracks and joints in complex scenarios and proposes an improved model, ViR-Deeplabv3+, based on DeepLabv3+. First, the model replaces the traditional backbone network with the Vision Transformer (ViT) with global perception ability. This enables the model to no longer be limited to extracting local information when processing image data, but to capture the global context features of the image more efficiently, thereby enhances subsequent segmentation tasks. Secondly, residual connections between the ViT and the Hollow Space Pyramid Pooling (ASPP) module are ingeniously introduced. The design concept of residual connection effectively solves problems such as the gradient vanishing problem, ensuring that the rich feature information from ViT can be smoothly and unobstructedly transmitted to the ASPP module for further fusion and mining of multi-scale features. Finally, we conducted model training and ablation experiments based on the self-built dataset (including crack and seam samples). The results showed that the mean intersection and union ratio (mIoU) of ViR-Deeplabv3+ reached 75.27%, which was 2.97% higher than that of the baseline model Deeplabv3+. This scheme provides an effective solution for precisely detecting and segmenting cracks and joints in complex scenarios, and has important practical application value.
Keywords: Image segmentation; Crack; Seam; ViT; Residual connection

1 INTRODUCTION

With the acceleration of urbanization and the aging of infrastructure, detecting cracks on the surface of concrete structures has become an important task to ensure the safety and durability of buildings. Cracks not only affect the load-bearing capacity of structures but may also cause secondary problems such as leakage and corrosion, posing a serious threat to public safety. Traditional crack detection methods mainly include visual inspection and manual measurement, which are simple and low-cost but are greatly influenced by human factors and difficult to monitor continuously [1]. In addition, manual inspection has limitations such as low efficiency, strong subjectivity, and high cost. In practical scenarios, the similarity in appearance between cracks and prefabricated joints, complex background interference, and the scarcity of datasets pose severe challenges to the generalization ability and practicality of models. Therefore, research on effective detection and segmentation of cracks and joints is crucial for ensuring the safety and reliability of infrastructure such as bridges, roads, and buildings.

Over the past few decades, crack detection has been continuously carried out and has achieved significant accomplishments. In the research methods based on object detection for crack detection, Pratibha et al. [2] deployed an automated process based on a deep learning object detection model, YOLOv5, By capturing and accurately locating cracks in masonry structures through bounding boxes, the training time of the model is relatively short and can be used for real-time crack detection; Marin B et al. [3] proposed a new detection method, progressive detection, which adopts the architecture of Faster R-CNN object detector to provide crack detection in images. From the perspective of detection, they re examined the binary classification of images with and without cracks, minimizing the crack loss rate to the greatest extent possible; Wang et al. [4] proposed an improved method based on the SSD algorithm, adjusting the combination of the number of prior boxes at different resolutions in the original SSD algorithm to achieve high-precision crack recognition for images with noise. In the research methods based on image segmentation for crack detection, Lau et al. [5] proposed a U-Net-based network architecture that replaces the encoder with a pre-trained ResNet-34 neural network and uses a "single cycle" training plan based on cyclic learning rates to accelerate convergence. Their model achieved higher F1 scores on CFD datasets compared to other models; Attard et al. [6] demonstrated that Mask R-CNN can be used to localize cracks on concrete surfaces and obtain their corresponding masks to aid extract other properties that are useful for inspection; Yao et al. [7] added an RFB multi-branch convolution module to the Deeplabv3+ model [8], replaced the backbone of Deeplabv3+ with Mobilenetv2, and replaced all ordinary convolutions in the algorithm with depthwise separable convolutions, improving the segmentation accuracy and detection efficiency of the Deeplabv3+ model for bridge cracks.

Most of these works focus on a single category of cracks and lack joint segmentation and geometric parameter calculation for cracks and joints. In actual engineering, the joints of precast concrete slabs are highly similar in morphology to real cracks. However, most current models only perform single-category detection for cracks without considering the interference of joints, leading to an increase in misjudgment rates. False positives in crack detection, such as misidentifying construction joints as cracks, waste resources, and delay critical repairs. [9]. Moreover, public datasets typically only contain crack samples and lack images simultaneously labeled with both cracks and joints, which limits the models' ability to distinguish between the two. Research shows that when the test set includes joints, the average mIoU of existing models drops by approximately 12% [10]. Although current research has provided valuable insights and techniques in the field of crack detection, the models' ability to segment cracks and joints remains to be improved when dealing with the highly similar morphologies of the two.

To address the aforementioned issues, we propose a ViR-deeplabv3+ model, which integrates Vision Transformer (ViT) and residual connections to improve the deeplabv3+ model for image segmentation of cracks and joints. The main contributions of this paper are as follows:

(1) Replace the backbone network Xception of DeepLabv3+ with Vision Transformer (ViT), and utilize its self-attention mechanism to capture global context dependencies, thereby overcoming the limitations of traditional convolutional networks in long-distance feature modeling.
(2) Introducing residual connections between the ViT and ASPP modules alleviates the vanishing gradient problem in deep networks, enhances the multi-scale feature fusion capability, and improves the edge segmentation accuracy of cracks and joints.
(3) By integrating publicly available data with self-collected data, a concrete structure image dataset containing annotations of cracks and joints is constructed. The sample size is effectively expanded through data augmentation techniques to enhance the generalization ability of the model.
(4) The ablation experiments verified the effectiveness of ViT and residual modules. The mIoU of ViR-Deeplabv3+ was significantly improved compared to the baseline model, and it demonstrated stronger robustness under complex background interference.

2 METHOD

The Deeplabv3+ model is a powerful semantic segmentation framework. Its classic version uses Xception as the backbone network and combines Atrous Spatial Pyramid Pooling (ASPP) with a decoder structure to achieve multi-scale feature fusion and fine edge recovery. However, because of the limitations of Xception in extracting global context dependencies and the potential problems, such as gradient vanishing when training deep networks, this study proposes the ViT-deeplabv3+ model. By replacing the original Xception with ViT (Vision Transformer) as the backbone network and introducing a residual connection module between the backbone and ASPP, the feature transmission efficiency and semantic expression ability are enhanced. The overall structure is shown in Figure 1.

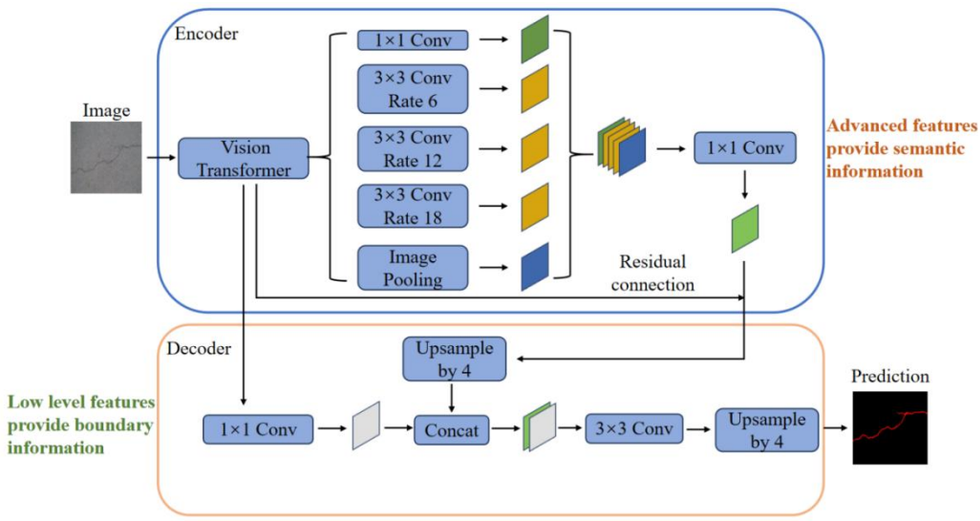


Figure 1 Overall Structure of ViT-deeplabv3+

2.1 ViT Feature Extraction Module

The Vision Transformer (ViT) is an image classification network based on the Transformer architecture, as shown in Figure 2. It divides the image into fixed-size patches and flattens them to be processed by the Transformer. ViT employs the self-attention mechanism to capture the global context information of the image, thereby demonstrating stronger performance than traditional convolutional neural networks (CNNs) in many computer vision tasks.

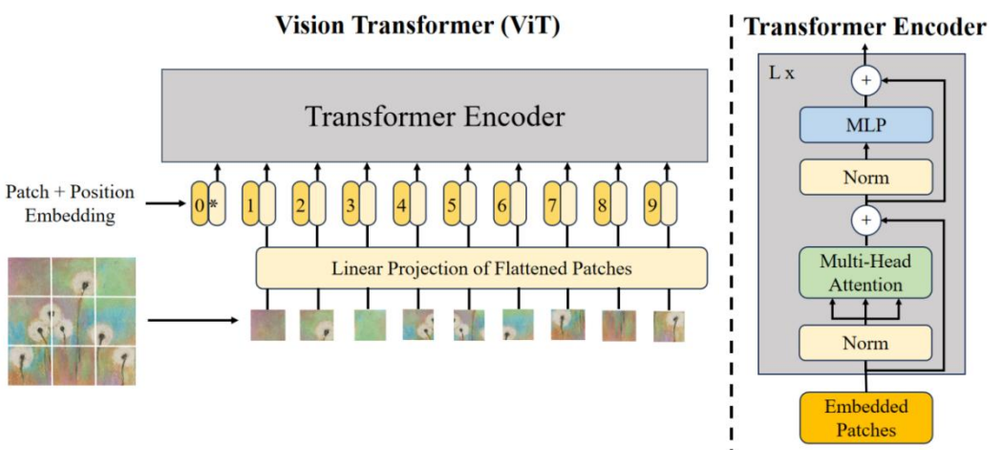


Figure 2 Network Structure of Vision Transformer

For the input image $I \in R^{H\times W\times C}$, it is first divided into $N = \frac{H\times W}{P^2}$ parts. A $P \times P$ small block, each small block through. Linearly embed the mapping into a high-dimensional space to form the input features:

$$z_i = E \cdot \text{Flatten}(I_i) + e_i \quad (1)$$

Here, E is the linear embedding matrix, $\text{Flatten}(I_i)$ represents the flattening operation of the -th small block, e_i is the position encoding used to retain the spatial position of the small block in the original image, and z_i is the feature of the -th block. In this way, ViT can encode the spatial information of the image into a sequence of inputs for the Transformer to process.

In ViT, the core computational module is the self-attention mechanism (Self-Attention). The self-attention mechanism assigns an attention weight to each input by computing the relationships among Query, Key, and Value, thereby performing a weighted sum of different parts of the input sequence. The calculation formula for self-attention is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Among them, Q is the query matrix, K means the key matrix, V indicates the value matrix, and d_k denotes the dimension of the key. Under the multi-head self-attention mechanism, multiple attention heads calculate in parallel and concatenate the results to provide richer context information.

By stacking multiple Transformer encoder layers, ViT can capture multi-level context information from local to global, which makes it more flexible and efficient than traditional CNNs in handling global dependencies. After being processed by multiple self-attention layers, the feature map output by ViT will be used as the input for the subsequent Deeplabv3+ model.

2.2 Residual Connection

Deep neural networks may encounter problems of vanishing or exploding gradients during training, especially when there are many layers, which can lead to unstable training. Residual connection is an effective solution. It forms a shortcut path by directly adding the input to the output, thereby avoiding the problem of vanishing gradients. Specifically, the mathematical representation of the residual connection is:

$$y = \mathcal{F}(x, W_i) + x \quad (3)$$

Among this is the input, $\mathcal{F}(x, W_i)$represents the output after a series of operations (such as convolution, activation, etc.), and y means the final output.

In this study, residual connections are introduced between the ViT backbone network and the ASPP module. Specifically, the feature maps output by ViT are added to the multi-scale feature maps processed by the ASPP module, thereby enhancing information flow and alleviating the vanishing gradient problem in deep networks. This process can be expressed as:

$$F_{out} = \text{ASPP}(F_{ViT}) + F_{ViT} \quad (4)$$

Among them, F_{ViT} is the high-level feature extracted by ViT, and F_{out} denotes the output processed by ASPP. The final feature after the residual connection is used as the output of the model.

By introducing residual connections, the network can more efficiently propagate gradients, thereby facilitating the learning of deeper features. Additionally, residual connections help preserve high-level semantic information, enabling the network to better retain detailed information during multi-scale feature fusion, and improving edge recovery and segmentation accuracy.

In summary, we propose an improved Deeplabv3+ model that uses ViT as the backbone network to overcome the limitations of traditional convolutional networks (such as Xception) in handling global context information. Additionally, we introduce a residual connection module between ViT and the ASPP module to address the gradient vanishing problem that may occur during the training of deep networks. Experimental results show that the Deeplabv3+ model with ViT as the backbone network, combined with residual connections, demonstrates better performance in semantic segmentation tasks.

3 EXPERIMENTAL EVALUATION

3.1 Dataset Construction

In this study, we integrated publicly available crack datasets with our own collected data to construct a specialized dataset of concrete crack and joint images. This dataset contains 151 high-resolution images, including 118 crack samples and 33 seam samples. All images were collected from diverse real-world engineering scenarios (as shown in Figures 3 and 4), covering various environmental conditions and structural types to ensure the representativeness and generalization ability of the data. In the data preprocessing stage, a standardized process was adopted: first, all original images were uniformly adjusted to a resolution of 513×513 pixels; then, the labelme annotation tool was used to conduct meticulous manual annotation on the self-collected data, automatically generating corresponding JSON format annotation files (as shown in Figure 5); finally, these JSON files were converted into annotation masks suitable for image segmentation tasks.

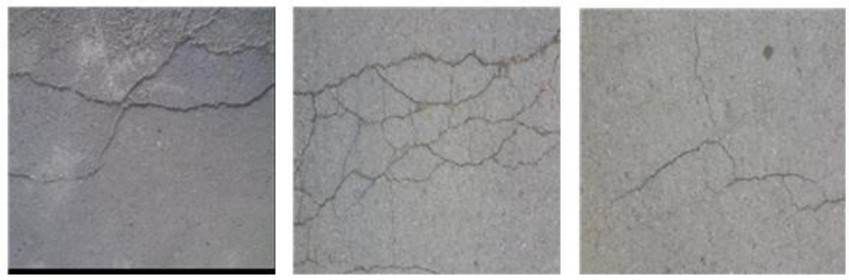


Figure 3 Partial Crack Images

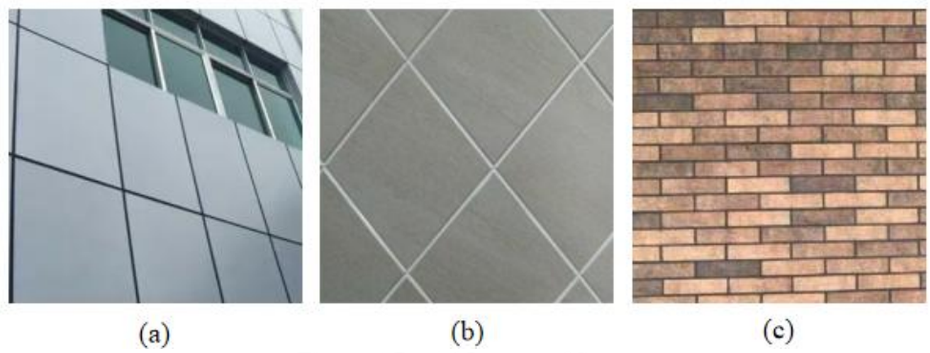


Figure 4 Partial Images of Seams

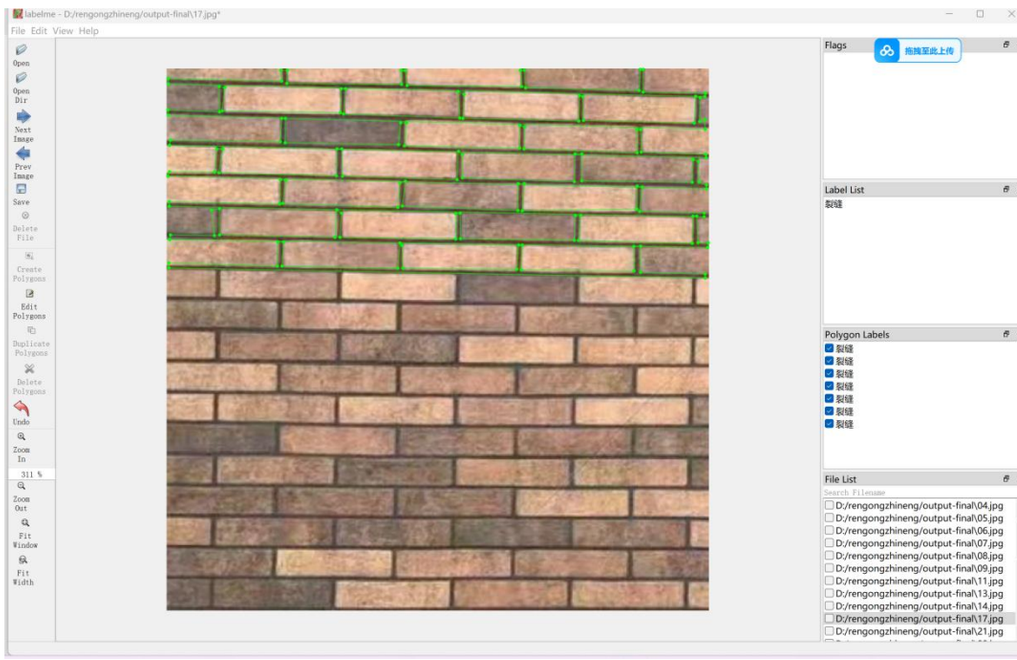


Figure 5 LabelMe Annotated Seam Image

Besides, to optimize the data quality, we applied Gaussian filtering for noise reduction to all images. For the issue of insufficient sample size, to effectively expand the dataset, data augmentation techniques were adopted. For crack images, two random augmentation methods were each applied twice, resulting in 354 augmented samples (118×3). For seam images, each of the two augmentation methods was applied nine times, ultimately yielding 330 augmented samples (33×10). (The specific augmentation effects are shown in Figures 6 and 7.) Through this strategy, not only was the data scale significantly increased, but also the diversity of key features and the consistency of annotations in the samples were ensured.

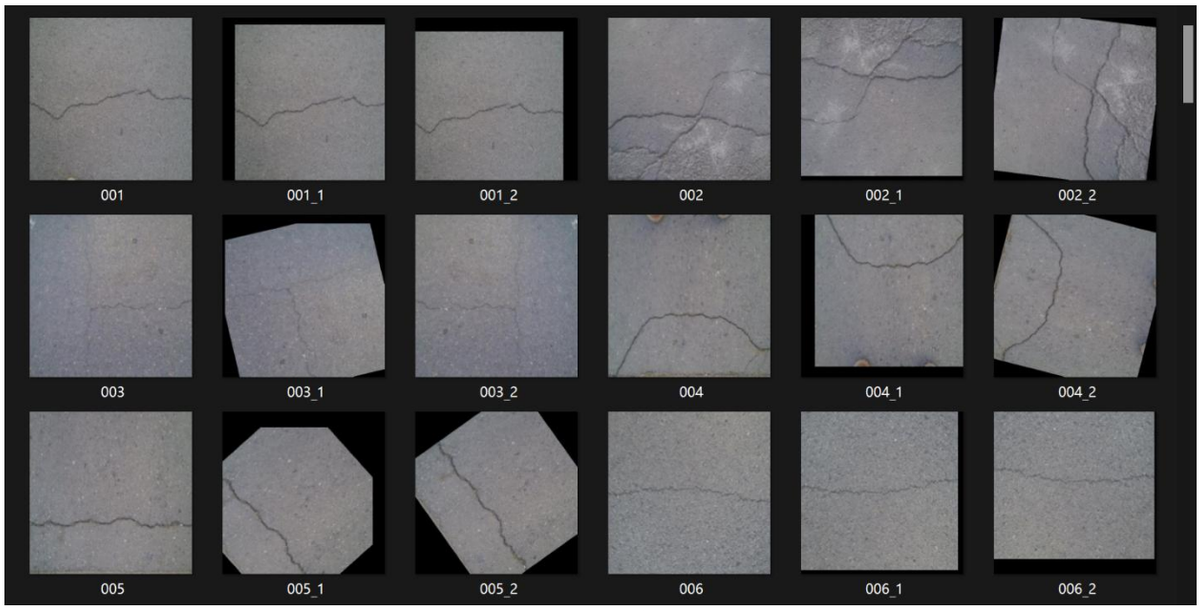


Figure 6 Cracks Image after Preprocessing

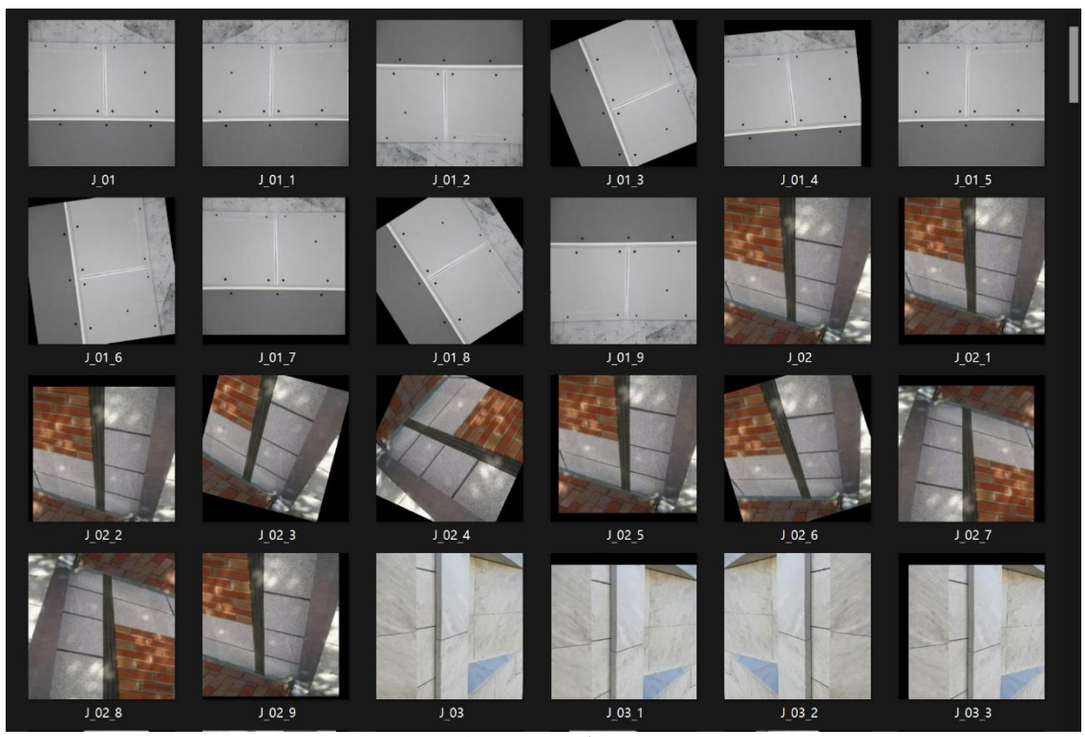


Figure 7 Preprocessed Seam Image

3.2 Model Training and Its Analysis

The experimental environment of this paper is RTX 3090 (24GB) GPU, PyTorch 2.0.0, Python 3.10 (Ubuntu 22.04), and Cuda 12.4. During the experiments, each data domain was divided into a training set and a validation set in an 8:2 ratio. The Adam optimizer was used with an initial learning rate of 0.1, which decreased stepwise as the training epochs increased. We set the batch size to 16 and use the mean Intersection over Union (mIoU) as the evaluation metric. Both the proposed ViR-Deeplabv3+ and the comparison method Deeplabv3+ were trained for 100 epochs under the same experimental settings. The experimental results show that the proposed ViR-Deeplabv3+ achieved the best mIOU on the test set, as detailed in Table 1.

Table 1 Comparison Results of Different Backbones

Method	Backbone	Segmentation accuracy (mIoU) (%)
Deeplabv3+	Xception	72.3
ViR-Deeplabv3+	Vision Transformer	75.27

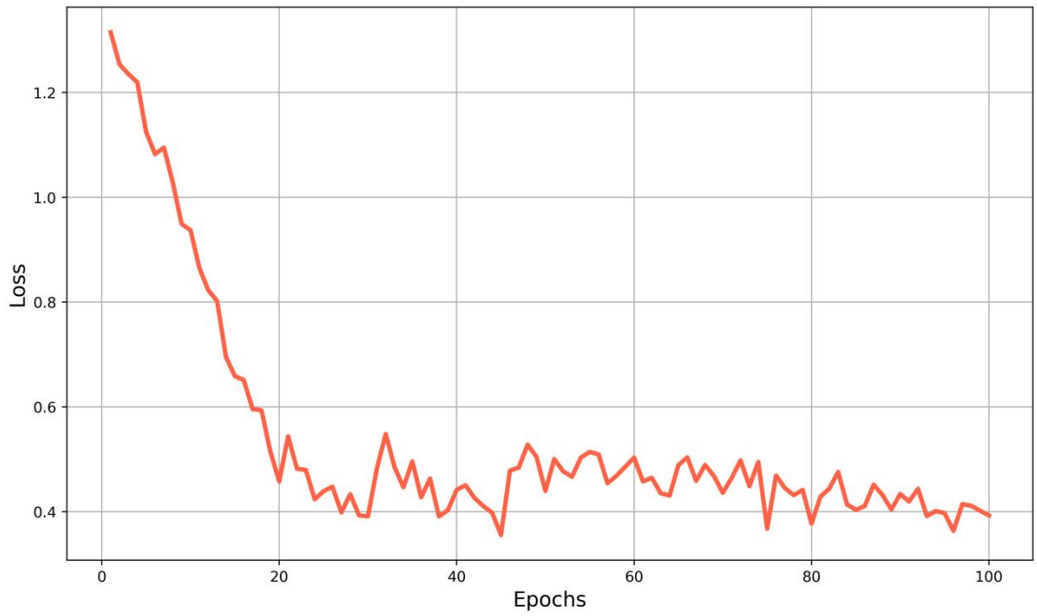
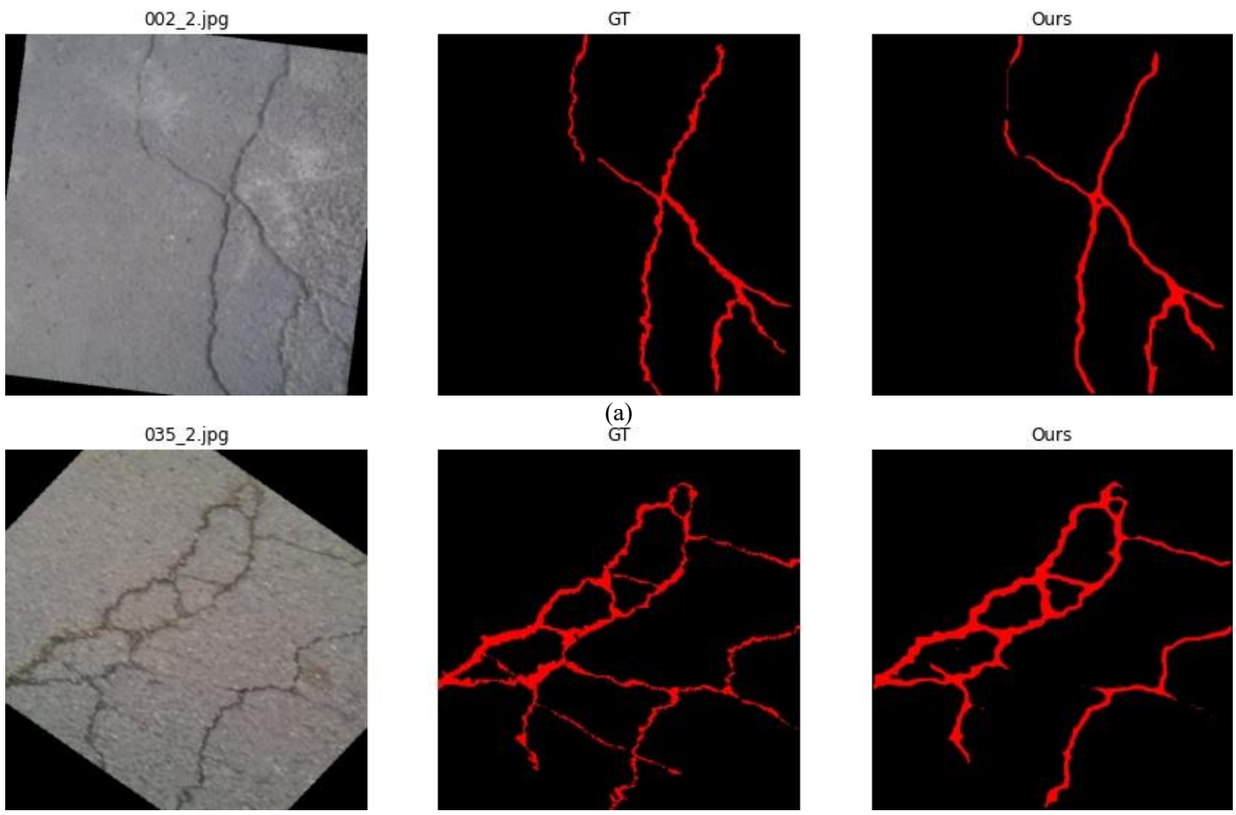


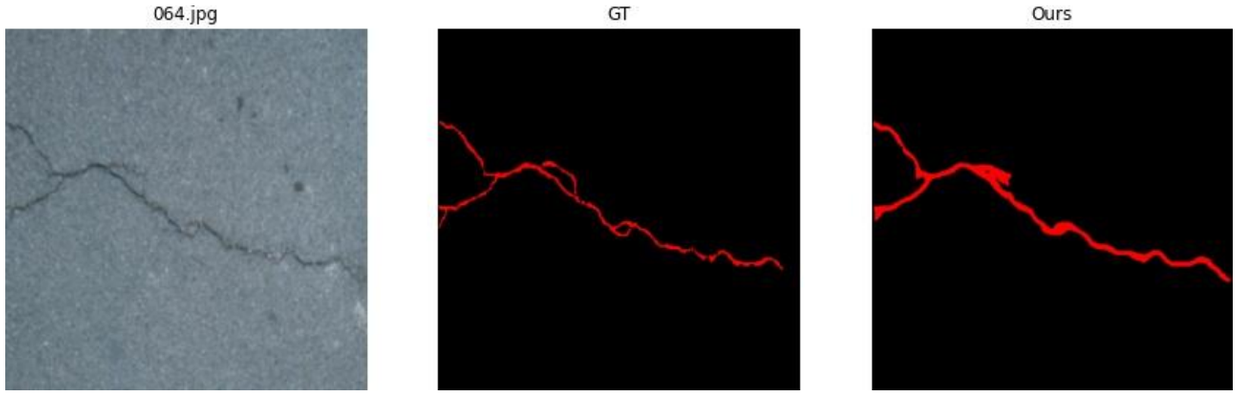
Figure 8 Model Training Loss Results

As shown in Figure 8, after 100 training epochs, the training loss of the deep learning model decreased from the initial value of 2.8 to 0.4. The convergence curve is smooth without obvious oscillations, indicating that the model has effectively learned the segmentation features of cracks and joints on the training set. By comparing the performance differences between the Vision Transformer and Xception backbones (see Table 1), the results show that Vision Transformer performs better in segmentation tasks: its deep residual structure, through skip connections, it alleviates the vanishing gradient problem and can accurately capture the slender morphological features of cracks and the regular edge information of joints. However, although Xception reduces the computational load through depthwise separable convolution, its ability to capture local details is significantly weakened under complex background interference (such as concrete surface texture and stains), resulting in limited segmentation accuracy, as shown in Figures 9 and 10. Experiments have demonstrated that the multi-scale feature fusion mechanism of Vision Transformer is effective in distinguishing morphologies. Similar cracks and joints play a crucial role.



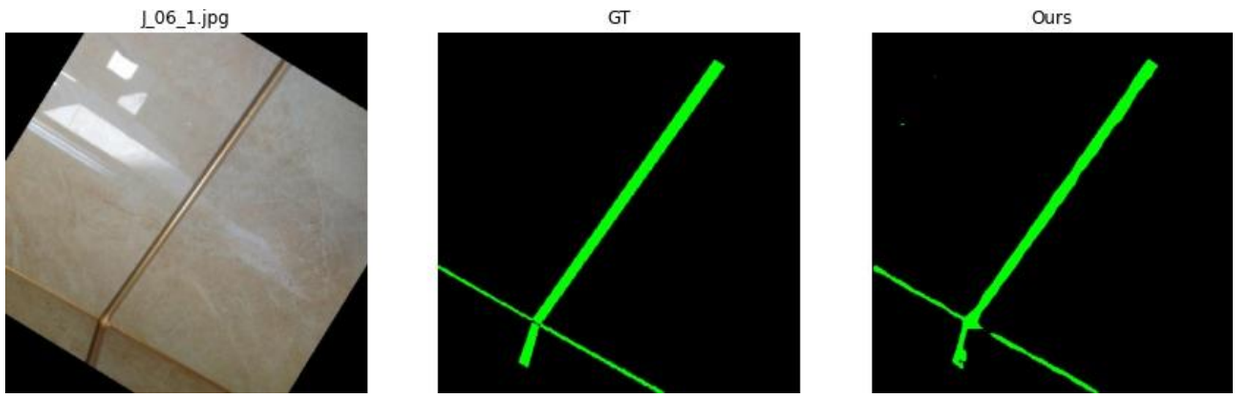
(a)

(b)

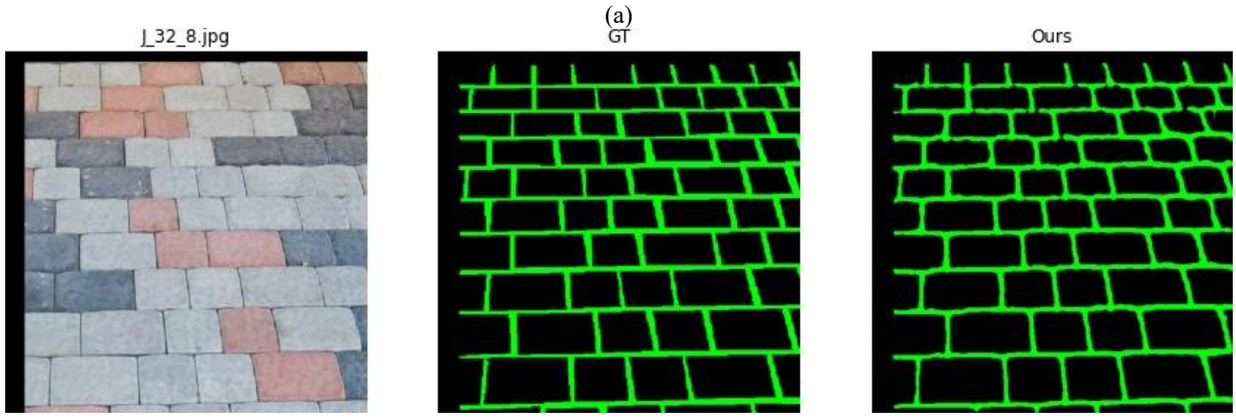


(c)

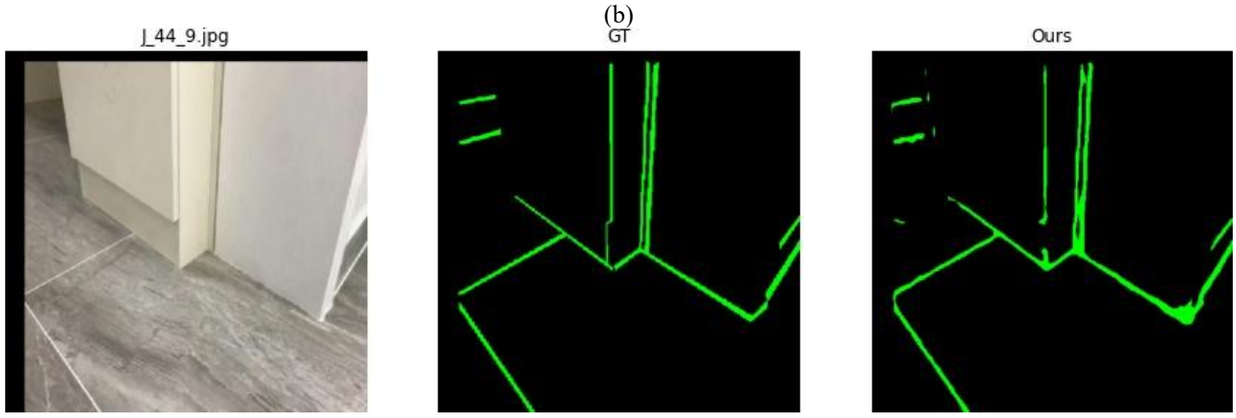
Figure 9 Model Training Results of Cracks



(a)



(b)



(c)

Figure 10 Model Prediction Results of the Joint Seam

3.3 Ablation Experiment

To verify the effectiveness of the proposed method, ablation experiments were conducted, and the specific results are shown in Table 2. By comparing the results of ViR-Deeplabv3+ (w/o ViT (use Xception)) and ViR-Deeplabv3+, it can be seen that when ViT is used to replace Xception, the segmentation accuracy (mIoU) increases from 73.3% to 75.27%, indicating that the introduction of ViT significantly improves the segmentation performance. This may be attributed to ViT's stronger ability to capture global information and its advantages in handling objects of different scales and structures. Meanwhile, by comparing the results of ViR-Deeplabv3+ (w/o residual) and ViR-Deeplabv3+, it is found that after adding the residual module, mIoU increases from 74.3% to 75.27%, suggesting that the residual module also contributes to improving the segmentation accuracy. It can alleviate the gradient vanishing problem in deep network training and facilitate cross-layer information transmission, enabling the model to better integrate feature information

from different levels. In conclusion, both the ViT and residual module in the ViR - ViR-ViR-Deeplabv3+ method contribute to enhancing the segmentation accuracy. The synergy of these components enables the model to achieve higher accuracy in semantic segmentation tasks, validating the effectiveness of the proposed method.

Table 2 Average Intersection over Union of Ablation Experiments under Different Networks

Method	Segmentation accuracy (mIoU) (%)
ViR-Deeplabv3+(w/o ViT (use Xception)	73.3
ViR-Deeplabv3+(w/o residual)	74.3
ViR-Deeplabv3+	75.27

4 CONCLUSIONS AND OUTLOOKS

The ViR-Deeplabv3+ model proposed in this paper significantly improves the segmentation accuracy of cracks and joints through the collaborative optimization of the ViT backbone network and residual connections, solving the misjudgment problem caused by traditional models' neglect of joint interference. Experiments show that the improved model achieves an mIoU of 75.27% on the self-built dataset, a performance improvement of 2.97% compared to the original DeepLabv3+ (with Xception backbone), verifying the effectiveness of the global modeling ability of ViT and residual connections. Besides, the constructed specialized dataset provides a data foundation for the joint segmentation research of cracks and joints.

Unfortunately, due to the scarcity of datasets, our dataset only contains images of either seams or cracks, but not both types simultaneously. In the future, we will further optimize the model's performance and practicality, expand data diversity, collect more samples of mixed cracks and seams in complex scenarios, and enhance the model's environmental adaptability. Additionally, we will deploy the model in actual engineering scenarios such as bridges and roads and conduct long-term stability tests to verify its robustness and generalization ability, promoting the transformation of intelligent detection technology from theoretical research to engineering application.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Ma Jing, Liu Lin, Lv Suyan. Monitoring and evaluation technology of cracks in concrete bridges. Comprehensive Corrosion Control, 2025, 39(4): 185-188.

[2] Pratibha K, Mishra M, Ramana G V, et al. Deep learning-based yolo network model for detecting surface cracks during structural health monitoring//International Conference on Structural Analysis of Historical Constructions. Cham: Springer Nature Switzerland, 2023, 179-187.

[3] Marin B, Brown K, Erden M S, et al. Automated masonry crack detection with faster R-CNN//2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), Lyon, France, 2021, 333-340. DOI: 10.1109/CASE49439.2021.9551683.

[4] Wang Yanhua, He Junze, Zhang Mingzhou, et al. Complex-environment concrete crack recognition based on SSD and pruned neural network. Journal of Southeast University (English Edition), 2023, 39(4): 393-399.

[5] Lau Stephen L H, Chong Edwin K P, Yang Xu, et al. Automated pavement crack segmentation using U-Net-based convolutional neural network. IEEE Access, 2020, 8, 114892-114899.

[6] Attard L, Debono C L, Valentino G, et al. Automatic crack detection using mask R-CNN[C]//2019 11th international symposium on image and signal processing and analysis (ISPA), Dubrovnik, Croatia, 2019, 152-157. DOI: 10.1109/ISPA.2019.8868619.

[7] Yao Yukai, Guo Baoyun, Li Cailin, et al. Bridge crack segmentation algorithm based on improved Deeplabv3+. Journal of Shandong University of Technology (Natural Science Edition), 2024, 38(2): 21-26.

[8] Chen Liang-Chieh, Zhu Yukun, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. 2018. DOI: https://doi.org/10.48550/arXiv.1802.02611.

[9] Torres-Acosta A A, Martínez-Madrid M. Residual life of corroding reinforced concrete structures in marine environment. Journal of Materials in Civil Engineering, 2003, 15(4): 344-353.

[10] WANG Y, ZHANG H. Impact of joint-crack mixed datasets on semantic segmentation models. IEEE Transactions on Image Processing, 2022, 31(5): 2345-2356.